Introduction — Reductive and Nonreductive Physicalism

A Short Survey of Six Decades of Philosophical Discussion Including an Attempt to Formulate a Version of Physicalism Against the Background of C. D. Broad's Concept of Emergence

Ansgar Beckermann

Materialism, or physicalism as it is now often called, is one of the main positions with regard to the mind-body problem. What, however, is the exact content of this position? What must one subscribe to if one wants to be a materialist or physicalist? The easy answer to these questions seems to be that someone is a materialist or physicalist if and only if he claims that there is nothing but physical objects and events, and that all that can be meaningfully said about these objects and events can be expressed in physical language. A claim like this, however, would seem to involve the denial not only of the existence of any nonmaterial mental substances, but also the denial of any mental events and properties. That is to say, if materialism were to be understood in this way the only possible kind of materialism would be eliminative materialism.

But most materialists and physicalists today would refuse to be counted among the adherents of eliminative materialism. So, if we want to avoid the conclusion that the views of these philosophers are altogether incoherent, there must be a different explication of what is meant by materialism or physicalism, an explication which, on the one hand, allows the existence of mental events and properties, but which, on the other hand, claims that these mental events and properties are derivative from, or dependent on, physical events and properties in such a way that it is legitimate to view the realm of the physical as the basic realm, the realm on which everything else is dependent. In other words, if someone wants to hold a position that is materialistic, but not eliminativistic, that is to say, if someone wants to be a materialist without denying the existence of mental events and properties, he has to claim that there is a special relation between the mental and the physical, a relation that makes it appropriate to call the

realm of the physical 'basic' or 'fundamental' or 'that which underlies everything else'. But what could that relation be? The classical answer to this question is: Reduction or reducibility. Someone is a materialist or physicalist if he claims that mental phenomena — although there really are such phenomena — are nonetheless *reducible* to physical phenomena. If this is true, nonreductive physicalism, however, seems to be a *contradictio in adjecto*. So, it behooves us to take a short look at the development of the debate on the concept of physicalism in order to see what reasons have led so many philosophers to adopt this very view and why they think that nonreductive physicalism is yet a materialist position.

1. Semantic Physicalism

When, at the beginning of the thirties, physicalism became one of the main tenets of most members of the Vienna Circle, it was not in the first place meant to be a theory about the relations between the mental and the physical. It was set forth as a theory of science and especially as a theory of the foundations of science. Its main claim was that "physical language" is "the universal language of science" and that therefore, even protocol sentences, i. e., those basic sentences of science which form the starting point of all hypotheses and the evidence by which to check their validity, should be couched not in phenomenalistic, but in physical language. The main reason for this claim was that only physical language is intersubjective and therefore apt to provide a suitable basis for scientific research. It is easy to see, however, that this very general theory also had to have important consequences for the status of psychology and sociology. If physical language is the universal language of science, psychology and sociology can be sciences only insofar as they can be reformulated in terms of physical language.[1]

When Carnap became convinced by the arguments of Neurath and others, he himself formulated the principles of what then became the mainstream within logical empiricism in two seminal papers "Die physikalische Sprache als Universalsprache der Wissenschaft" und "Psychologie in physikalischer Sprache" both of which appeared in ERKENNTNIS in

[1] It is not entirely clear, however, how this term is to be understood. Sometimes physical language is equated with what often is called "thing language". So understood, talk of rocks, mountains, and chairs qualifies as talk in physical language. On other occasions it seems that the descriptive vocabulary of physical language comprises only terms referring to physical quantities that can be ascribed to space-time-points. With regard to Carnap's own proposal (2) (cf. below, pp. 3—4) it seems that, at least in this context, he is taking physical language to be thing language.

1932. As the title already suggests, the first of these papers was concerned with the general thesis that physical language should be universal, i.e., the language in which all scientific knowledge should be expressed. The second paper was devoted to the more special claim that even psychology could and should be rephrased in this way. Right at the beginning Carnap writes:

> In what follows we intend to explain and to establish the thesis that *every sentence of psychology may be formulated in physical language*. To express this in the material mode of speech: *all sentences of psychology describe physical occurrences, namely, the physical behavior of humans and other animals*. This is a sub-thesis of the general thesis of *physicalism* to the effect that *physical language is a universal language*, that is, a language into which every sentence may be translated. (p. 165)

The material mode of speech, however, may be misleading. That is to say, physicalism, according to Carnap, ought not to be understood as requiring the sentences of psychology to concern themselves only with physical events and situations.

> The thesis, rather, is that psychology may deal with whatever it pleases, it may formulate its sentences as it pleases – these sentences will, in every case, be translatable into physical language. (p. 166)

Physicalism for Carnap therefore entails the thesis that each sentence S of psychology can be translated into physical language, i.e., that for each such sentence there exists a physical sentence S' such that S and S' have the same meaning. Carnap himself adduces as an example the sentence

(1) Mr. A is now excited

which, he claims, can be translated into the physical sentence

(2) Mr. A's body (especially his central nervous system) is now in a state that is characterized by a high pulse and rate of breathing, which, on the application of certain stimuli, may even be made higher, by vehement and factually unsatisfactory answers to questions, by the occurrence of agitated movements on the application of certain stimuli, etc.[2]

That these two sentences indeed have the same meaning is, according to Carnap, shown by the fact that "every protocol sentence which confirms [(1)] also confirms [(2)] and *vice versa*" (p. 166).

The physicalism that Carnap and other members of the Vienna Circle propounded can be called *semantic* physicalism since it is possible to formulate its main thesis as follows:

---

[2] Cf. Carnap (1932 b, pp. 170–172).

(SP) Any sentence S of any science, and *a fortiori* any sentence S couched in psychological terms, can be translated into a sentence S' of physical language that has the same meaning as S.[3]

This version of physicalism, however, is hardly plausible as can already be shown by reference to Carnap's own example. If we have a closer look at sentence (2), there are at least three remarkable shortcomings of this alleged translation of sentence (1). First, (2) contains two occurrences of the expression "on the application of *certain* stimuli" which are no more than placeholders for descriptions of the exact stimuli that are pertinent in this context. As long as these expressions are not replaced by specific physical descriptions of the stimuli envisaged, (2) can at best be viewed as a translation scheme but not as a complete translation of sentence (1).[4]

The second and in a way related problem of the purported translation is marked by the "etc." at the end of sentence (2). This expression may look innocent, but it in fact hides a serious problem, namely the problem that it is hard to tell which physical conditions or behavioral dispositions should be included in the translation of sentence (1). That Carnap puts an "etc." at the end of (2) shows that he himself did not think the list explicitly given in (2) to be complete. But how shall we go about filling the gap? Is there any clear criterion that tells us what to include and what not? And, finally and most important, what reason do we have for believing that there really is a finite set of conditions which will do the job?

Very similar considerations have been set forth in the debate on what could be called the law of practical syllogism

(3) If x wants p, then x will do A if x believes that A is a suitable means for achieving p.

Everyone now admits that (3) as it stands cannot be counted as true, since there are a number of conditions which, when fulfilled, have the effect that x will not do A even if he wants p and believes that A is a suitable means for achieving p. But if one tries to remedy this weakness by adding further conditions like "if there is nothing that x wants more than p and that is incompatible with his doing A", "if x is able to do A", "if x is not too tired", etc., there always seem to remain counterexamples which show that someone might refrain from doing A even if all the original and all these additional conditions are fulfilled, i. e., that show that even

[3] Given this definition, it is obvious that logical behaviorism is just one version of semantic physicalism.

[4] One might argue that the term "certain" does not function as a placeholder, but as an existential quantifier in this context. I think, however, that this is not what Carnap had in mind.

the amended version of (3) does not hold in every case.[5] And, what is perhaps even more important, there seems to be no hope that this situation can be overcome, since, although (3) has been under discussion for at least thirty years, it has not been possible to find a list of conditions that copes with all these counterexamples. For similar reasons it seems very improbable that Carnap's proposal (2) should be more successful in this respect.

The third and perhaps decisive problem of the translation of (1) by (2), however, is constituted by the fact that contrary to what Carnap suggests, (2) is certainly *not* a sentence of physical language. This in fact is quite obvious, since (2) contains expressions like "unsatisfactory answers to questions". And it is not only hard to see how the terms "question" and "answer" could be given a purely physical interpretation, but the expression "*unsatisfactory* answer" obviously cannot be translated into physical language at all.

Even if this would turn out, however, to be a problem only of Carnap's particular example, there are more general reasons to be very suspicious of the possibility of translating sentences couched in a mental vocabulary into sentences of physical language that do not themselves contain any mental terms. Think again of the concept of wanting, or to be more precise, of the sentence

(4) x wants p.

How could we translate this sentence into physical language?

Inspired by the law of practical syllogism, one could try to start with the following definition hoping that the remaining mental terms contained in the definiens will themselves prove translatable into physical language one by one.[6]

(5) x wants p iff
x does A if x believes that A is a suitable means for achieving p, if there is nothing that x wants more than p and that is incompatible with his doing A, if x is able to do A and if x is not too tired.

---

[5] This kind of argument is well known from philosophical discussions on the problem of action explanation. For an especially lucid exposition cf. Lanz (1987; forthcoming).

[6] One might argue that this proposal is doomed to fail from the very beginning, since it is flatly circular in virtue of containing "wants" in the definiens. Against this objection one could, however, reply that the expressions "x wants p" and "x wants p more than q" are at least syntactically different and that, moreover, it might turn out to be easier to find a translation into physical language for the latter than for the former.

Besides the problem that it is quite unclear whether the conditions mentioned in the definiens add up to something that is indeed sufficient for doing A if one wants p, it, however, turns out to be more than difficult to cash out on the mentioned hope. How, e. g., shall we translate

(6) x believes that A is a suitable means for achieving p?

Thinking in terms of behavioral dispositions, something like this seems at least roughly correct:

(7) If x wants p, then x does A.

Even if this were the case however, (7) is of no help since it contains the term "wanting" which we originally tried to translate into physical language. But there seems to be no other practicable way to translate or paraphrase (6). In short: Every time we try to explicate the meaning of a mental expression in terms of behavioral dispositions we find ourselves in the situation that we cannot formulate the conditions of the disposition except by using other mental expressions. And if we try to explicate the meaning of these expressions, we have to use the mental expressions we wanted to get rid of in the first place. There seems to be no way out of this circle.

2. *Identity Theory*

So, semantic physicalism was bound to fail. But it was not before the end of the fifties that a promising alternative was propounded: the identity theory that was put forward, amongst others, by U. T. Place and J. J. C. Smart.[7] Exploiting Frege's distinction between meaning and reference, Place and Smart argued that from the fact that mental terms cannot be translated into physical terms it does *not* follow that the referents of these terms must be different. Even if the expression "x has pain" does not have the same meaning as the expression "x's C-fibers are firing", pain may still be *identical* with the neural process of C-fiber activation since it might well be that the two expressions, though not synonymous, *de facto* refer to the same kind of processes or states. Whether they actually do is an empirical question that has to be decided by further research. The thesis of the identity theory can therefore be stated like this:

(IT) Mental phenomena are identical with brain states in the sense that for each mental predicate M there exists a neurophysiological

[7] The two seminal papers were Place (1956) and Smart (1959). Almost at the same time Feigl developed a very similar theory in Feigl (1957).

> predicate P such that, though M is not synonymous with P, M *de facto* denotes the same kind of brain states as P.

This formulation is however not entirely correct. For I already mentioned that Place and Smart actually do not claim that mental phenomena are identical with brain states but only that they well might be. That they really are is an hypothesis which will be confirmed or refuted by empirical investigation.

There appear to be numerous identities of this kind, i. e., statements of the form "A is B" that are true despite the fact that the terms "A" and "B" are not synonymous. Water is identical with H_2O, lightning with an electrical discharge, and the temperature of an ideal gas with the mean kinetic energy of its molecules.[8] It is however interesting to ask what are the reasons because of which we believe that they are true. Take, e. g., the third of the mentioned examples. What are the reasons which lead us to believe it to be true that the temperature of an ideal gas is identical with the mean kinetic energy of its molecules? The classic answer to this question is that the truth of this identity statement follows from the fact that classical thermodynamics is reducible to statistical mechanics, where this in turn means that the laws of classical thermodynamics can be deduced from the laws of statistical mechanics with the aid of suitable bridge laws.[9] From statistical mechanics we can, e. g., deduce the law

(8) $\qquad M * V = 2/3 * \mathrm{N} * E,$

where N is the number of the molecules of the gas in the volume V, M the average of the momenta transferred from the molecules to the walls of the container, and E the mean kinetic energy of the molecules. And from this law we can deduce Boyle-Charles's law

(9) $\qquad p * V = \mathrm{N} * k * T$

by means of the bridge laws

(10) $\qquad p = M$

(11) $\qquad 2/3 * E = k * T.$

---

[8] Smart calls identity statements like "Water is identical with H_2O" *contingent statements* (cf. Smart 1959, p. 58). This, however, does not mean that he claims that the identities expressed by these statements are contingent (instead of necessary in Kripke's sense), but only that the statements themselves are contingent, and not analytic (i. e., that they are not true in virtue of the meanings of terms "water" and "H_2O" alone).

[9] I am not convinced that this classical Nagelian account is an adequate explication of what is meant by reduction, i. e., I much prefer the account that was set forth in Hooker (1981) and applied to the mind-body problem in P. M. Churchland (1985) and P. S. Churchland (1986). Cf. my contribution to this volume pp. 107—109.

These bridge laws are necessary in the deduction of (9) since (9) is couched in terms that do not even occur in statistical mechanics. Usually bridge laws are taken to have the form of nomological biconditionals which connect one term of the reducing theory with a corresponding term of the reduced theory or of equations which do the same job for a whole family of terms.

If one accepts this classical account, the consequences for the identity theory are straightforward. Mental phenomena are identical with states of or processes in the brain if and only if psychology is reducible to neurobiology. That is to say, if there are bridge laws that connect each mental predicate with a suitable neurobiological predicate in such a way that the laws of psychology can be deduced from neurobiology with the aid of these bridge laws. The existence of nomological biconditionals connecting mental with neurobiological predicates or mental with neural properties is therefore a necessary condition if mental phenomena are to be identical with neural phenomena although it is by itself not yet sufficient to establish this identity.

Obviously, the identity theory of Place and Smart has the great advantage of showing that there is a feasible version of materialism or physicalism which need not presuppose that mental expressions can be translated into physical ones. But even according to the identity theory, mental phenomena are reducible to neural phenomena in the sense that there are bridge laws which show that each mental predicate is necessarily coextensive with some physical predicate. Since there are however good arguments against such reducibility, many philosophers started to try and find out whether there are other options, i. e., versions of physicalism that do not even presuppose reducibility. Before following these attempts in more detail, we should first have a closer look at the arguments that have been put forward against the reducibility of the mental.

A first line of argument, the functionalist line,[10] suggests itself if we ask once more what the identity theory actually amounts to. According to this theory, mental phenomena are identical with neural phenomena if the laws of psychology can be deduced from neurobiology with the aid of suitable bridge laws. But what form will such a deduction probably take? There is an easy answer to this question. Let us assume that $P_1, \ldots, P_n$ are the neural properties that are piece by piece paired with the mental properties $M_1, \ldots, M_n$ by the bridge laws in question. Then obviously, the laws of psychology can be deduced from neurobiology with the aid of these bridge laws if for every law L of psychology the image law L' —

[10] This line of argument was especially set forth by Putnam (1960; 1967 a; 1967 b) and Fodor (1968).

i. e., the law that results if the psychological predicates occurring in L are replaced by their neurobiological counterparts — can be deduced from neurobiology. If for example

(12) For all x: if $M_1(x)$ and $M_7(x)$, then x does A

is a psychological law, (12) can be deduced from neurobiology and the bridge laws if it is possible to deduce from neurobiology the image law

(12′) For all x: if $P_1(x)$ and $P_7(x)$, then x does A.

For in this case (12) follows from (12′) and from the pertinent bridge laws simply by substitution of equivalents.

What shall we say however if there are two individuals a and b such that for a the mental properties $M_1, \ldots, M_n$ can successfully be paired with the neural properties $P_1, \ldots, P_n$ while for b the same mental properties can equally successfully be paired with quite another set of, say, electro-mechanical properties $Q_1, \ldots, Q_n$? In other words, what shall we say if we can deduce from neurobiology the lawlike sentence

(12″) If $P_1(a)$ and $P_7(a)$, then a does A

and from the theory of electronics the lawlike sentence

(12‴) If $Q_1(b)$ and $Q_7(b)$, then b does A?

And if the same can be done not only for the law (12) but for all the laws of psychology? Two answers seem inevitable. First, obviously a and b both instantiate the laws of psychology. So, there is no reason for assuming that a does have certain mental properties while b does not. Both of them have the same right to be ascribed the same mental predicates. Psychologically or mentally, there is no difference between them. But, and this is the second answer, in the case described it is also obvious that the mental properties $M_1, \ldots, M_n$ can be *identical* neither with the neural properties $P_1, \ldots, P_n$ nor with the electro-mechanical properties $Q_1, \ldots, Q_n$, since b may have any of the mental properties $M_1, \ldots, M_n$ without having any of the neural properties $P_1, \ldots, P_n$ as well as a may have any of these mental properties without having any of the electro-mechanical properties $Q_1, \ldots, Q_n$. This is reflected by the fact that in this case actually none of the necessary bridge laws

(13) For all x: $M_i(x)$ if and only if $P_i(x)$

or

(14) For all x: $M_i(x)$ if and only if $Q_i(x)$

holds.

On the other hand, there certainly is a relation that holds between the mental properties M_1, ..., M_n and the neural properties P_1, ..., P_n in *a* and the electro-mechanical properties Q_1, ..., Q_n in *b*, respectively. But this is not the relation of identity, but that of *realization*. In *a*, the mental properties M_n, ..., M_n are realized by the neural properties P_1, ..., P_n, because these neural properties bear exactly those relations to each other and to the (perceptual) inputs and the (behavioral) outputs of *a* that are characteristic of the corresponding mental properties, i. e., exactly those relations that are expressed by the laws of psychology. And for just the same reason the mental properties M_1, ..., M_n are realized in *b* by the electro-mechanical properties Q_1, ..., Q_n. So, the functionalist line of argument against the identity theory can be summarized like this: Mental properties cannot be identical with any physical properties whatsoever because in different individuals they can be realized in multiple ways by quite different sets of properties.[11]

A second line of argument against the identity theory of Place and Smart is the line of argument that D. Davidson set forth in favour of his famous doctrine of anomalous monism. Anomalous monism is the claim that each particular mental event is identical with some physical event, although there are no type identities between the mental and the physical since mental properties are irreducible, by law or definition, to physical properties.[12] In the present context it is the second part of this theory that is more important. So, let us concentrate on the reasons Davidson offers for his claim that there cannot be any type identities between mental and physical properties. As is quite clear from his writings his main reason for claiming this is that there can be no strict psychophysical laws and therefore *a fortiori* no psychophysical bridge laws connecting mental and physical properties in a nomological way. Why is this? According to Davidson, there can be no strict psychological laws because mental and physical predicates are incommensurable in much the same way as the predicates "is an emerald" and "is grue"[13]. For "laws" couched in terms of incommensurable predicates, such as the "law"

---

[11] In (1981 b, pp. 8–9) Fodor claims that it is not multiple realizability that is decisive. In his view, the real point is that even if there were bridge laws connecting each mental predicate "M_i" with a corresponding physical predicate "P_i" it would be, to say the least, highly improbable that the "P_i" were physical kind predicates, i. e., predicates denoting natural kinds in physics.

[12] To say that mental properties are irreducible, by law or definition, to physical properties, in this context is just to say that there are no corresponding biconditionals, holding either by law of nature or in virtue of the meaning of the relevant predicates.

[13] "is grue" is true of an object x if and only if x has been observed before time t and x is green or if x is blue. Cf. Goodman (1954).

(15) All emeralds are grue,

cannot be nomological, i. e., they cannot support subjunctive counterfactuals, they are not confirmable by empirical evidence, etc.[14] The reasons for the incommensurability of mental and physical predicates however, may be different from the reasons of the incommensurability of predicates like "is an emerald" and "is grue". For in the former case, incommensurability is mainly due to the fact that the ascription of mental properties is to a high degree constrained by assumptions of rationality that are incompatible with the constitutive principles which guide the ascription of physical properties.

3. Supervenience

That there can be no type identities between mental and physical properties does not in Davidson's view preclude the possibility "that mental characteristics are in some sense dependent, or supervenient, on physical characteristics" in the sense "that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect" (1970, p. 214). For many philosophers supervenience therefore became the idea they pinned their hopes on. For supervenience, as it seemed to them, was the most promising concept in terms of which a viable theory of *nonreductive* physicalism could be formulated.[15]

There are however reasons to be sceptical whether the concept of supervenience really is suited for this job.[16] First, there is the problem of how this concept is exactly to be understood. To be sure, the intuitive idea behind it can be easily stated:

(S) A set A of properties supervenes on a set B of properties if and only if any two objects that are indiscernible with respect to their B-properties are also indiscernible with respect to their A-prop-

[14] Obviously, Goodman thought that (15) is not lawlike because "is grue" is not a projectible predicate. According to Davidson, however, it is not because of the properties of single predicates that (15) is not lawlike, but because of the fact "that the predicates 'is an emerald' and 'is grue' are not suited to one another". (1970, p. 218) In his view, the sentence "All emerires are grue" may be as lawlike as "All emeralds are green". (Here "x is an emerire" is supposed to be true of an object x if and only if x has been observed before time t and x is an emerald or if x is a sapphire.)

[15] To the following remarks cf. the section on supervenience of my contribution to this volume (below pp. 94—100) and Kim (1984; 1990).

[16] Cf. esp. Grimes (1988) and Kim (1989; 1990).

erties, i. e., if any two objects that differ with respect to some A-property also differ with respect to at least one B-property.

But there are different ways to give this idea a precise formulation, depending on what modal force one wants to associate with it. If one thinks supervenience to imply only that the principle (S) holds within any possible world, one gets what has been called *weak supervenience*:

(WS) For any possible world w and for any objects x and y, if x has in w the same B-properties that y has in w, then x has in w the same A-properties that y has in w.

If, on the other hand, one thinks that supervenience implies that the principle (S) holds across possible worlds, one gets the corresponding concept of *strong supervenience*:

(SS) For any possible worlds w_i and w_j and for any objects x and y, if x has in w_i the same B-properties that y has in w_j, then x has in w_i the same A-properties that y has in w_j.

And if, to mention a third possibility, one thinks that supervenience does not depend so much on the B-indiscernibility of particular objects as on the B-indiscernibility of whole worlds, one gets what has been called *global supervenience*:

(GS) Any two worlds w_i and w_j which are indiscernible with respect to B-properties are also indiscernible with respect to A-properties.

Weak and global supervenience however are much too weak relations to be suitable foundations for versions of materialism or physicalism that really deserve the name. Materialism or physicalism implies at least that all mental properties are determined by or dependent on physical properties. But there can be no relation of determination or dependency that does not support corresponding counterfactuals. That is to say, if A is determined by, or dependent on, B, then the counterfactual "If B were the case, then so would A" must be true. Weak supervenience however, only implies that all objects that have the same B-properties *within* a possible world w also have the same A-properties *in that world*. So, if we assume that an object a has the A-property F in w and if we abbreviate the conjunction of all the B-properties a has in w by G, then weak supervenience only implies that the universal conditional "For all x, if $G(x)$, then $F(x)$" is true in w. But this is compatible with the assumption that even in the nearest possible world w', a lacks F, though it has G. And if this is so, the counterfactual "If a were G, it would also be F" is false even in w.

Global supervenience, on the other hand, is compatible with there being

> a world which differs physically from this world in some most trifling respect (say, Saturn's rings in that world contain one more ammonia molecule) but which is entirely devoid of consciousness, or has a radically different, perhaps totally irregular, distribution of mental characteristics over its inhabitants (say, creatures with brains have no mentality while rocks are conscious). (Kim 1987, p. 321)

So, global supervenience is, after all, compatible with the assumption that there may be two objects *a* and *b* in two possible worlds *w* and *w'* which share all their physical properties, but differ in their mental properties as radically as you like. And this seems to be entirely incompatible with what we have in mind when we say that the mental is determined wholly by the physical.

Strong supervenience, therefore, seems to be the only option left. But even this is no longer undisputed.[17] For, first, dependency relations are usually understood to be asymmetric whereas the relation of strong supervenience by itself is neither symmetric nor asymmetric. If A strongly supervenes on B this does neither imply nor preclude that B also strongly supervenes on A. In (1990) Kim gives a nice example in which strong supervenience fails to be asymmetric:

> ... think of a domain of perfect spheres. The surface area of each sphere strongly covaries with its volume, and conversely, the volume with the surface area. And we don't want to say either determines, or depends on, the other, in any sense of these terms that implies asymmetry. There is only a functional determination, and dependence, both ways; but we would hesitate to impute a metaphysical or ontological dependence either way. (p. 13)

But what is more, and this is the second problem, even if A strongly supervenes on B in a way that is *de facto* asymmetric, this does not show that A is determined by, or dependent on, B since there might be a third set of properties on which both A and B depend, so that the strong supervenience of A on B is only a side effect of the more fundamental relations between A and C and B and C, respectively. Again, Kim gives a nice example:

> As a possible example consider this: I've heard that there is a correlation between intelligence as measured by the IQ test and manual dexterity. It is possible that both manual dexterity and intelligence depend on certain genetic and developmental factors, and that intelligence strongly covaries with manual dexterity but not conversely. If such were the case, we would not consider intelligence to be dependent on, or determined by, manual dexterity. (p. 15)

We, therefore, must conclude that what all the different types of supervenience provide us with is nothing more than as many different types of

[17] Cf. esp. Grimes (1988), Kim (1990).

property covariation. None of them implies that mental properties really depend on physical properties.

And this is not all that can be put forward against the idea of supervenience as the foundation of a plausible formulation of physicalism. Even if, as we might assume for the sake of argument, strong supervenience would imply that mental properties depend – say, in a causal way – on physical properties, this would not suffice to make strong supervenience a relation suited for a plausible formulation of physicalism. For even then, strong supervenience would be compatible with such dualistic positions as psychophysical parallelism or epiphenomenalism. What supervenience, at best, provides us with is the truth of some nomological universal conditionals or biconditionals of the form "For all x, if $G(x)$, then $F(x)$" or "For all x, $G(x)$ if and only if $F(x)$", where F and G are suitable mental and physical predicates. But the truth of such nomological universal conditionals or biconditionals is compatible with a number of dualistic positions. So, the concept of supervenience is not of much help if one wants to formulate a viable version of nonreductive physicalism, since physicalism demands more than just causal or nomological dependence. If supervenience was the only hope for nonreductive physicalism, this kind of physicalism would indeed have no good prospects.

4. Physicalism as the Theory that Mental Properties are not Emergent

The crucial question, therefore, is whether there are further alternatives in the offing. As already stressed, to show that mental properties are nomologically correlated with, or even causally dependent on, physical properties, is not enough if one wants to show that physicalism is true. But what more is needed? What relation must hold between the physical and the mental properties or states of a person for physicalism to be true? One might be tempted to say that what is essential is that individuals have their mental properties *because of* or *in virtue of* their physical properties[18] where these expressions are understood in a noncausal sense. But this is of no help unless one can give a satisfactory account of what is meant by "because of" and "in virtue of" in the present context. Is there any sense of these terms that does not rely on property identities? Exactly at this point it proves to be very helpful to have a retrospect on the debate concerning the concept of emergence that was of great importance in the first half of this century. This will become obvious especially when we have a closer look at C. D. Broad's contribution to this debate.

[18] Cf. Kim (1990, p. 16).

At the beginning of the twenties Samuel Alexander and C. Lloyd Morgan used G. H. Lewes' distinction between merely "resultant" and truly "emergent" properties[19] to formulate a theory which became known by the title "Emergent Evolutionism".[20] According to this theory the whole universe develops in such a way that the configurations of its basic material elements become more and more complex. This growth in complexity, however, is not a gradual process exhaustively describable in purely quantitative terms. For when the complexity of material configurations reaches a certain critical level, genuinely novel properties emerge, properties that have never been instantiated before. This evolutionary process is moreover hierarchically structured. For complex objects with their emergent properties may also combine to form still more complex entities so that further novel properties emerge. According to Samuel Alexander there are four major stages to be distinguished in the evolution of the universe: first, the emergence of matter out of space and time; second, the emergence of life out of complex configurations of matter; third, the emergence of consciousness and mentality out of biological processes; and fourth, the emergence of deity out of consciousness. This may sound odd, and it is not very surprising that C. Lloyd Morgan had difficulties with Alexander's first and last stages of the process of emergent evolution. But this is not of much importance in the present context: What is important is the notion of an emergent property that plays such a central role in Alexander's account. That is to say, the notion of a property of a complex object that is not only novel in the sense that it has never been instantiated before, but also in the sense that it could not even have been predicted in advance.

C. D. Broad was the first to give a proper explication of the concept of an emergent property. He, like many of his contemporaries, used the concept of emergence to formulate a third way between vitalist and mechanistic theories for the explanation of the phenomenon of life – or to be more precise, of the characteristic behavior of living beings. Vitalists claimed that everything that is characteristic of living beings – breathing as well as digestion, procreation as well as goal directed and intelligent behavior – can only be explained by the assumption that there is a certain nonmaterial component, a substance-like entelechy or elan vital, which is present in all beings that exhibit the characteristic features of life, and absent in all nonliving entities. Mechanists, on the other hand, claimed that even the phenomena of life can be explained completely by the kind

[19] Cf. Lewes (1875, pp. 412 ff.).

[20] To the following exposition of the main claims of this doctrine cf. Kim this volume, pp. 121–122.

and arrangement of the material parts of living beings and the general laws of nature which hold for these parts as well as for any other objects in the world.

In order to find a third way between vitalism and mechanism Broad first tried to give a taxonomy of the kinds of theories that can be advanced if one is concerned with the explanation of the characteristic behavior of certain kinds of things or systems. In this context he first proposed to distinguish between *component* theories on the one hand and what might be called structural theories on the other hand. The characteristic feature of component theories is the claim that the behavior of a certain class of objects is, in part, to be explained by "the presence of a peculiar *component* which does not occur in anything that does not behave in this way" (1925, p. 55). Vitalism, of course, is a component theory. But it is not due to this fact that vitalism is mostly regarded as an odd and unscientific theory, but rather to the fact that the specific components postulated, entelechies or elans vitales, are of a completely different kind than normal material entities and that, moreover, they are entirely hypothetical entities which have so far eluded all the usual empirical methods of natural science.

The rivals of component theories, structural theories as I have called them, concur in denying

> that there need be any peculiar *component* which is present in all things that behave in a certain way and is absent from all things which do not behave in this way. [They both say] that the components may be exactly alike in both cases, and [try] to explain the difference of behaviour wholly in terms of difference of structure. (Broad 1925, pp. 58—59)

But although all structural theories share this very general point of view, one has to distinguish between two types of such theories which differ radically in their view of the laws which connect the properties of the components with the characteristic behavior of the complex wholes which these components make up. These two types of theories Broad calls *emergent* theories and *mechanistic* theories. Emergent theories claim that

> the characteristic behaviour of the whole *could* not, even in theory, be deduced from the most complete knowledge of the behaviour of its components, taken separately or in other combinations, and of their proportions and arrangements in this whole. (p. 59)

According to mechanistic theories on the other hand

> the characteristic behaviour of the whole is not only completely *determined by* the nature and arrangement of its components; in addition to this it is held that the behaviour of the whole could, in theory at least, be *deduced* from a sufficient knowledge of how the components behave in isolation or in other wholes of simpler kind. (p. 59)

What is essential for emergent and mechanistic theories is summarized by Broad as follows:

> Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A, B, and C in a relation R to each other; that all wholes composed of constituents of the same kind as A, B, and C in relations of the same kind as R have certain characteristic properties; that A, B, and C are capable of occurring in other kinds of complex where the relation is not of the same kind as R; and that the characteristic properties of the whole R(A, B, C) cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B, and C in isolation or in other wholes which are not of the form R(A, B, C). The mechanistic theory rejects the last clause of this assertion. (p. 61)

The core of Broad's analysis of the notion of an emergent property can be formulated like this:

(E) A property F of a system S, made up of the constituents $C_1, \ldots, C_n$ standing in a certain relation R to each other, is emergent if and only if (a) there is a law to the effect that all systems with the same make-up have F, and if (b) F nonetheless cannot, even in theory, be deduced from the most complete knowledge of the properties of the components $C_1, \ldots, C_n$ in isolation or in systems with a different make-up.

It is not easy to see how the term "deduce" is to be understood in this context and why Broad uses such a complicated formulation as "from the most complete knowledge of the properties of the components $C_1, \ldots, C_n$ in isolation or in systems with a different make-up". I take it however that Broad's definition (E) can be rephrased like this:

(E′) A property F of a system S, made up of the constituents $C_1, \ldots, C_n$ standing in a certain relation R to each other, is emergent if and only if (a) there is a law to the effect that all systems with the same make-up have F, and if (b) it nonetheless cannot be proved from the general laws of natural science which apply to all kinds of objects and not only to objects like $C_1, \ldots, C_n$ standing in relation R to each other, that systems which have the same make-up as S have all the features (or exhibit exactly that behavior) which are (is) characteristic of property F.[21]

At least, (E′) seems to me to come very close to what Broad probably had in mind. If this is true there is however a close connection between

[21] For a more detailed analysis of Broad's definition of the concept of emergent properties cf. the section on emergence in my contribution to this volume (below pp. 100–106).

the concept of emergence and the concept of realization that was first elaborated in the context of functionalism.[22] If we have a set of functional states F_1, ..., F_n, then each of these states is characterized by the (causal) relations it bears to the other members of the set and to the inputs and outputs of the system it is a state of. In a given system S, the functional state F_i therefore can be said to be realized by the physical state G_i if and only if there is a set of physical states G_1, ..., G_n of system S such that these states bear exactly those relations to each other and to the inputs and outputs of S that are characteristic of the states of the set F_1, ..., F_n and if G_i in this context corresponds to F_i. Now, as has already been said, the relations that the members of F_1, ..., F_n bear to each other and to the inputs and outputs of a system are exactly what makes them the kind of states they are, i. e., these relations are their characteristic features. Generalizing from the case of functionalism, the relation of realization can therefore be defined like this:

(R) If a system S is in a (mental) state F at time t, F can be said to be realized at t by the (physical) state G if and only if S is in G at t and G has in S at t all the (monadic, relational, etc.) features that are characteristic for states of kind F.

So defined, the concept of realization is the exact complement of the concept of emergence. A state or property F of a system S is emergent if it is not realized by a physical state of S. Or what amounts to the same thing: if it is realized by a physical state of S, it is not emergent.[23]

In the present context the distinction between emergent and nonemergent properties or the distinction between properties that are realized by physical states of the systems in which they are instantiated and properties that are not is of great importance. For it shows that even if a macroproperty F is nomologically dependent on the physical states of the system in which it is instantiated, there is an important distinction to be made: F may or may not be "deducible from the most complete knowledge of the properties of these states", i. e., F may or may not be realized by the states it depends on. So, it might be a good idea to use this distinction to formulate a viable version of physicalism in the following way.

---

[22] Cf. above pp. 8–10.

[23] I thus think that Kim is wrong in claiming that the "realization" relation is the converse of the "emergence" relation, i. e., that "P realizes M" is true if and only if "M is emergent from P" is true (cf. his contribution to this volume p. 133). The reason for this is simply that, in my view, the relation of realization is a much stronger relation than Kim believes it to be.

(P) Mental states and properties are not emergent since they are always realized by the physical states of the systems they are states or properties of.

Should physicalism, understood in this way, i. e., as the claim that mental phenomena are not emergent since each mental state is realized by some physical state, count as reductive or nonreductive? In my view, the answer to this is yes and no. On the one hand, this kind of physicalism is as reductive as it can be. It claims that individuals possess whatever mental properties they have in virtue of their physical properties, that each mental state is realized by some physical state. And this, in a sense, implies that there are only physical states and nothing else. According to this kind of physicalism, there are no mental states above or in addition to the physical states of an individual. It is only that some of these physical states are also mental states.

On the other hand, physicalism in the sense of (P) does not imply that all that can be meaningfully said about an individual can be said in physical terms. Quite on the contrary, physicalism in the sense of (P) is born out of the insight that individuals which do not share any of their physical properties may nonetheless behave quite similarly and that these individuals may therefore be subsumed under the same mental concepts, despite the fact that they do not have anything in common that can be expressed in physical terms. This kind of physicalism therefore is all too willing to acknowledge that there are regularities of behavior that can only be formulated in mental terms.

But isn't that contradictory? How can it be that the mental vocabulary is indispensable when, in a sense, there are no mental states or properties that could be expressed by its concepts? As I see it, these questions rest on a wrong premiss. They presuppose that a concept can only apply to the elements of a set A if there is a common property, in the robust sense,[24] that is shared by all elements of A. But this need not be the case. Let, e. g., A be a set of objects that all behave similarly although this behavior is caused by quite distinct physical properties in the different elements of A. That all these elements behave in a similar way is reason enough to subsume them under the same concept. But from this it precisely does *not* follow that they must share a common property which is causally responsible for this behavior.

[24] As to the distinction between a pleonastic and a robust sense of the term "property" cf. Schiffer (1987). There is at least a strong resemblance between the analysis of physicalism given here and the view developed by Schiffer in (1987) and (1990). Cf. also Beckermann (forthcoming).

So, physicalism in the sense of (P) is reductive in that it implies that – in a sense – there is nothing but physical states. But it is also nonreductive in that it endorses the view that the mental vocabulary is nonetheless indispensable for the formulation of some regularities of behavior. Is this kind of physicalism, then, reductive enough to be a materialist position? In my view, the answer must be yes. For, on the one hand, its claim that all states are physical states certainly is a materialist claim, and, on the other hand, its concession of the indispensability of the mental vocabulary does not have any antimaterialist implications since it is fully compatible with the view that there are no mental properties, in the robust sense, that are expressed by the concepts of this vocabulary.[25]

References

Alexander, S. (1920) *Space, Time, and Deity. 2 Vols.* London: Macmillan.

Beckermann, A. (forthcoming) "States, State Types, and the Causation of Behavior". *Erkenntnis.*

Borst, C. V. (ed.) (1970) *The Mind-Brain Identity Theory.* London: Macmillan.

Broad, C. D. (1925) *The Mind and Its Place In Nature.* London: Routledge and Kegan Paul.

Carnap, R. (1932 a) "Die physikalische Sprache als Universalsprache der Wissenschaft". *Erkenntnis* 2, pp. 432–465. (The translation into English by M. Black appeared as monograph: Carnap, R., *The Unity of Science.* London 1934.)

– (1932 b) "Psychologie in physikalischer Sprache". *Erkenntnis* 3, pp. 107–142. (Translation into English in: Ayer, A. (ed.), *Logical Positivism.* New York 1959, pp. 165–198.)

Churchland, P. M. (1985) "Reduction, Qualia, and the Direct Introspection of Brain States". *Journal of Philosophy* 82, pp. 8–28.

Churchland, P. S. (1986) *Neurophilosophy: Toward a Unified Science of Mind/Brain.* Cambridge, Mass.: MIT Press.

Davidson, D. (1970) "Mental Events", in: Foster, L., and Swanson, J. W. (eds.), *Experience and Theory.* Amherst: University of Massachusetts Press, pp. 79–101. Reprinted in: Davidson, D., *Essays on Actions and Events.* Oxford 1980, pp. 207–225.

Feigl, H. (1958) "The 'Mental' and the 'Physical'", in: Feigl, H., Scriven, M., and Maxwell, G. (eds.), *Concepts, Theories, and the Mind-Body Problem. Minnesota Studies in the Philosophy of Science. Vol. 2.* Minneapolis: University of Minnesota Press, pp. 320–492.

Fodor, J. (1968) *Psychological Explanation.* New York: Random House.

– (1974) "Special Sciences, or The Disunity of Science as a Working Hypothesis". *Synthese* 28, pp. 97–115. Reprinted in: Fodor (1981 a), pp. 127–145.

– (1981 a) *Representations.* Cambridge, Mass.: MIT Press.

– (1981 b) "Introduction: Something on the State of the Art", in: Fodor (1981 a), pp. 1–31.

[25] I am very grateful to Jaegwon Kim for his helpful comments on an earlier draft of this introduction.

Goodman, N. (1954) *Fact, Fiction, and Forecast*. London: Athlone Press.
Grimes, Th. R. (1988) "The Myth of Supervenience". *Pacific Philosophical Quarterly* 69, pp. 152–160.
Hooker, C. A. (1981) "Towards a General Theory of Reduction". *Dialogue* 20, pp. 38–60, pp. 201–236, pp. 496–529.
Kim, J. (1984) "Concepts of Supervenience". *Philosophy and Phenomenological Research* 45, pp. 153–176.
– (1987) "'Strong' and 'Global' Supervenience Revisited". *Philosophy and Phenomenological Research* 48, pp. 315–326.
– (1989) "The Myth of Nonreductive Materialism". *Proceedings and Addresses of the American Philosophical Association* 63, pp. 31–47.
– (1990) "Supervenience as a Philosophical Concept". *Metaphilosophy* 21, pp. 1–27.
Lanz, P. (1987) *Menschliches Handeln zwischen Kausalität und Rationalität*. Frankfurt/M.: Athenäum.
– (forthcoming) "The Explanatory Force of Action Explanations", in: Bieri, P., and Stöcker, R. (eds.) *Reflecting Davidson*, Berlin/New York: Walter de Gruyter Verlag.
Lewes, G. H. (1875) *Problems of Life and Mind. Vol. 2*. London: Kegan Paul, Trench, Turbner, & Co.
Morgan, C. Lloyd (1923) *Emergent Evolution*. London: Williams & Norgate.
Place, U. T. (1956) "Is Consciousness a Brain Process?". *British Journal of Psychology* 47. Reprinted in: Borst (1970), pp. 42–51.
Putnam, H. (1960) "Minds and Machines", in: Hook, S. (ed.), *Dimensions of Mind*. New York, 138–164. Reprinted in: Putnam (1975), pp. 362–385.
– (1967 a) "The Mental Life of Some Machines", in: Castaneda, H. (ed.), *Intentionality, Mind, and Perception*. Detroit. Reprinted in: Putnam (1975), pp. 408–428.
– (1967 b) "Psychological Predicates", in: Capitan, W. H., and Merrill, D. D. (eds.) *Art, Mind, and Religion*. Pittsburgh. Under the title "The Nature of Mental States" reprinted in: Putnam (1975), pp. 429–440.
– (1975) *Mind, Language, and Reality. Philosophical Papers. Vol. 2*. Cambridge: Cambridge University Press.
Schiffer, St. (1987) *Remnants of Meaning*. Cambridge, Mass.: MIT-Press.
– (1990) "Physicalism". *Philosophical Perspectives* 4, pp. 153–185.
Smart, J. J. C. (1959) "Sensations and Brain Processes", in: *Philosophical Review* 58. Reprinted in: Borst (1970), pp. 52–66.